Estimation spline de quantiles conditionnels pour variables explicatives fonctionnelles

Spline estimation of conditional quantiles for functional covariates

Hervé CARDOT [a], Christophe CRAMBES [b] et Pascal SARDA [b,c]

[a] *Unité Biométrie et Intelligence Artificielle, INRA, Toulouse, BP 27, 31326 Castanet-Tolosan Cedex, France*
[b] *Laboratoire de Statistique et Probabilités, UMR CNRS C5583, Université Paul Sabatier, 31062 Toulouse Cedex, France*
[c] *GRIMM, EA 3686, Université Toulouse-le-Mirail, 31058 Toulouse Cedex, France*

Abstract

This Note deals with a linear model of regression on quantiles with the explanatory variable taking values in some functional space and a scalar response. We propose a spline estimator of the functional coefficient that minimizes a penalized L^1 type criterion (the penalization is of primary importance to get existence and convergence of the estimator), then we study the asymptotic behaviour of this estimator. *To cite this article: Hervé Cardot, Christophe Crambes, Pascal Sarda, C. R. Acad. Sci. Paris, Ser. I 336 (2003).*

Résumé

Cette Note a pour objet un modèle linéaire de régression linéaire sur quantiles lorsque la variable explicative est à valeurs dans un espace fonctionnel alors que la variable réponse est réelle. Nous proposons un estimateur spline du coefficient fonctionnel basé sur la minimisation d'un critère de type L^1 pénalisé (la pénalisation est primordiale pour avoir l'existence et la convergence de l'estimateur), puis nous étudions le comportement asymptotique de cet estimateur. *Pour citer cet article : Hervé Cardot, Christophe Crambes, Pascal Sarda, C. R. Acad. Sci. Paris, Ser. I 336 (2003).*

1. Introduction

Grâce aux performances accrues des appareils de mesure et à l'augmentation des capacités de stockage informatique, de nombreuses données sont collectées et sauvegardées sur des échelles temporelles ou des grilles spatiales de plus en plus fines (courbes d'évolution de températures, courbes spectrométriques, images satellites, ...). On est ainsi amené à traiter des données assimilables à des courbes ou plus

Email addresses: cardot@toulouse.inra.fr (Hervé CARDOT), crambes@cict.fr (Christophe CRAMBES), sarda@univ-tlse2.fr (Pascal SARDA).

généralement à des fonctions de variables continues (temps, espace). Ces observations sont appelées données fonctionnelles dans la littérature [8].

Des modèles de régression pour variables explicatives fonctionnelles ont été proposés, notamment lorsque la réponse est un scalaire : le modèle linéaire est introduit dans [7] tandis qu'un modèle fonctionnel non paramétrique est étudié dans [3]. Ces modèles sont construits dans le but d'estimer la moyenne conditionnelle. Cependant, dans certains cas, on s'intéresse plutôt à l'estimation de quantiles conditionnels, comme par exemple en agronomie (estimation de seuils de rendements), en médecine ou encore en fiabilité (voir par exemple [6]). Cette note est consacrée à l'étude de ce problème. Nous définissons dans la partie 2 un modèle linéaire de régression sur quantiles qui généralise au cadre fonctionnel le modèle proposé par Koenker et Bassett [5]. Nous proposons dans la partie 3 de construire un estimateur spline du paramètre fonctionnel du modèle. Des propriétés asymptotiques, dont l'obtention d'une borne supérieure pour la vitesse de convergence L^2 sont présentées dans la partie 4.

2. Modèle linéaire de régression fonctionnelle sur quantiles

Soit $(X_i, Y_i)_{i=1,\dots,n}$ un échantillon de couples indépendants de variables aléatoires définis sur un même espace de probabilités, de même loi que (X, Y). La variable explicative X est à valeurs dans un sous-espace fermé H de l'espace fonctionnel $L^2([0,1])$, l'espace des fonctions de carré intégrable sur l'intervalle $[0,1]$, et Y est la variable aléatoire réponse à valeurs dans $\mathbb{R}$.

Soit $\alpha \in]0,1[$ fixé ; en supposant que $\mathbb{E}|Y| < +\infty$, le quantile conditionnel d'ordre α de Y sachant $[X = x]$, $x \in H$, est le réel $g_\alpha(x)$ défini comme solution du problème de minimisation

$$\min_{a\in\mathbb{R}} \mathbb{E}(l_\alpha(Y-a)|X=x), \tag{1}$$

où $\mathbb{E}(.|X=x)$ désigne l'espérance conditionnellement à l'événement $[X=x]$ et l_α est la fonction définie par $l_\alpha(u) = |u| + (2\alpha-1)u$ (voir [5]).

Une généralisation directe du modèle de Koenker et Bassett [5] consiste à supposer que g_α est une forme linéaire continue définie sur H, c'est-à-dire qu'elle s'écrit :

$$g_\alpha(X) = \langle \Psi_\alpha, X\rangle = \int_0^1 \Psi_\alpha(t)X(t)\,dt, \tag{2}$$

la fonction Ψ_α appartenant à H et la notation $\langle .,. \rangle$ désignant le produit scalaire usuel de $L^2([0,1])$.

3. Construction d'un estimateur spline

L'estimateur défini ci-dessous repose sur une approximation du quantile conditionnel par une fonction spline. Il faut donc choisir une subdivision de l'intervalle $[0,1]$ en k sous-intervalles, $k = k_n \in \mathbb{N}^*$. Cette subdivision définit $k-1$ noeuds intérieurs que nous supposons équirépartis dans la suite. Les fonctions splines que nous considérons sont des polynômes de degré q par morceaux sur chaque sous-intervalle de $[0,1]$, et $(q-1)$ fois dérivables sur $[0,1]$, où $q \in \mathbb{N}$. L'espace de ces fonctions est un espace vectoriel de dimension $k+q$. Une base de cet espace vectoriel est l'ensemble des fonctions B-splines, que l'on notera $\mathbf{B_{k,q}} = {}^t(B_1, \dots, B_{k+q})$ (voir [2]).

On estime alors Ψ_α par une combinaison linéaire des B_l, $1 \le l \le k+q$: on est ramené à trouver un vecteur $\widehat{\boldsymbol{\theta}} = {}^t(\widehat{\theta}_1, \dots, \widehat{\theta}_{k+q}) \in \mathbb{R}^{k+q}$ tel que $\widehat{\Psi}_\alpha = {}^t\mathbf{B_{k,q}}\widehat{\boldsymbol{\theta}}$ avec $\widehat{\boldsymbol{\theta}}$ solution du problème de minimisation

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k+q}}\left\{\frac{1}{n}\sum_{i=1}^{n} l_\alpha\left(Y_i - \langle {}^t\mathbf{B_{k,q}}\boldsymbol{\theta}, X_i\rangle\right) + \rho \parallel ({}^t\mathbf{B_{k,q}}\boldsymbol{\theta})^{(m)} \parallel^2\right\}, \tag{3}$$

où $({}^t\mathbf{B_{k,q}}\boldsymbol{\theta})^{(m)}$ est la dérivée d'ordre m de ${}^t\mathbf{B_{k,q}}\boldsymbol{\theta}$, ρ est un paramètre de pénalisation pour contrôler le degré de "régularité" de l'estimateur cherché et $\|.\|$ est la norme associée au produit scalaire de $L^2([0,1])$.

Nous avons programmé cet estimateur dans le logiciel Splus en utilisant un algorithme basé sur la méthode des moindres carrés itérés repondérés (voir [6]). D'autres algorithmes assez courants dans la littérature portant sur les estimateurs de type L^1 pourraient être adaptés facilement à notre contexte.

4. Propriétés asymptotiques

Supposons que X soit du second ordre, c'est-à-dire que $\mathbb{E}\left(\|X\|^2\right) < +\infty$. Pour alléger les notations, nous supposons également dans cette partie que X est centrée ($\mathbb{E}(X) = 0$). On peut alors définir l'opérateur de covariance de X, opérateur Γ de $L^2([0,1])$ dans lui-même défini par $\Gamma h = \mathbb{E}(\langle X, h\rangle X)$. Cet opérateur induit une semi-norme définie par $\|h\|_2^2 = \langle \Gamma h, h\rangle$.

Afin d'établir le résultat de convergence pour l'estimateur $\widehat{\Psi}_\alpha$, on suppose que les hypothèses suivantes sont vérifiées.

$(H.1)$ $\parallel X \parallel \leq C_0 < +\infty, \quad ps.$

La fonction Ψ_α est supposée être p' fois dérivable et $\Psi_\alpha^{(p')}$ vérifie

$$\text{(H.2)} \qquad \left|\Psi_\alpha^{(p')}(t) - \Psi_\alpha^{(p')}(s)\right| \leq C_1|t-s|^\nu, \quad s,t \in [0,1],$$

où $C_1 > 0$ et $\nu \in [0,1]$. Dans la suite, on pose $p = p' + \nu$ et on suppose que $q \geq p \geq m$.

$(H.3)$ Les valeurs propres de Γ sont strictement positives.

$(H.4)$ Pour $x \in H$, la loi de Y conditionnellement à $[X = x]$ admet une densité f_Y^x continue et strictement positive au quantile d'ordre α.

Théorème 4.1 *Sous les hypothèses $(H.1) - (H.4)$ et si on suppose de plus qu'il existe $\beta, \gamma \in]0,1[$ tels que $k_n \sim n^\beta$ et $\rho \sim n^{(\gamma-1)/2}$, alors*

(i) $\widehat{\Psi}_\alpha$ existe sauf sur un ensemble dont la probabilité tend vers 0 lorsque n tend vers l'infini,

(ii) $\mathbb{E}\left(\parallel \widehat{\Psi}_\alpha - \Psi_\alpha \parallel_2^2 |X_1, \ldots, X_n\right) = O_P\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho} + \rho + \rho k_n^{2(m-p)}\right).$

Éléments de démonstration :

La démonstration reprend certains arguments de la preuve du résultat établi par He et Shi [4].

Le point (i) du théorème repose sur l'inversibilité de la matrice $\widehat{\mathbf{C}}_{\boldsymbol{\rho}} = \widehat{\mathbf{C}} + \rho\mathbf{G_k}$, $\widehat{\mathbf{C}}$ étant la matrice de terme général $\frac{1}{n}\sum_{i=1}^n \langle B_j, X_i\rangle\langle B_l, X_i\rangle$ et $\mathbf{G_k}$ la matrice de terme général $g_{jl} = \langle B_j^{(m)}, B_l^{(m)}\rangle$ pour $j, l = 1, \ldots k+q$. Or, il existe (voir [1]) une constante c telle que $\lim_{n\to+\infty} P\left(\left\{\omega/\lambda_{\min}(\widehat{\mathbf{C}}_{\boldsymbol{\rho}}) > c\frac{\rho}{k_n}\right\}\right) = 1$, ce qui entraîne le point (i).

L'hypothèse (H.2) implique qu'il existe une fonction spline, notée $\Psi_\alpha^\star = {}^t\mathbf{B_{k,q}}\boldsymbol{\theta}^\star$, vérifiant $\sup_{t\in[0,1]}|\Psi_\alpha^\star(t) - \Psi_\alpha(t)| \leq C_2/k_n^p$ (voir [2]). On obtient ainsi avec (H.1) : $\parallel \Psi_\alpha^\star - \Psi_\alpha \parallel_2^2 = O_P\left(1/k_n^{2p}\right)$. Il reste donc à étudier le terme $\mathbb{E}\left(\parallel \widehat{\Psi}_\alpha - \Psi_\alpha^\star \parallel_2^2 |X_1, \ldots, X_n\right)$. Pour cela, remarquons que pour $L > 0$ et $(\delta_n)_{n\in\mathbb{N}}$ une suite de réels positifs, on a :

$$P\left[\frac{1}{n}\sum_{i=1}^n \langle\widehat{\Psi}_\alpha - \Psi_\alpha^\star, X_i\rangle^2 + \rho\left\|(\widehat{\Psi}_\alpha - \Psi_\alpha^\star)^{(m)}\right\|^2 \leq L^2\frac{\delta_n^2}{n}|X_1, \ldots, X_n\right]$$

$$\geq P\left[\inf_{|\boldsymbol{\theta}|\geq L\delta_n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta}) > \sum_{i=1}^{n} f_i\left(\sqrt{n}\widehat{\mathbf{C}}_{\boldsymbol{\rho}}^{1/2}\widehat{\boldsymbol{\theta}} - \sqrt{n}\widehat{\mathbf{C}}_{\boldsymbol{\rho}}^{1/2}\boldsymbol{\theta}^{\star}\right)|X_1,\ldots,X_n\right],$$

où f_i est définie par :

$$f_i(\boldsymbol{\theta}) = l_\alpha\left[Y_i - \langle\, {}^t\mathbf{B}_{\mathbf{k},\mathbf{q}}\left(\frac{\widehat{\mathbf{C}}_{\boldsymbol{\rho}}^{-1/2}}{\sqrt{n}}\boldsymbol{\theta}+\boldsymbol{\theta}^{\star}\right), X_i\rangle\right] + \rho\left\|\left[{}^t\mathbf{B}_{\mathbf{k},\mathbf{q}}\left(\frac{\widehat{\mathbf{C}}_{\boldsymbol{\rho}}^{-1/2}}{\sqrt{n}}\boldsymbol{\theta}+\boldsymbol{\theta}^{\star}\right)\right]^{(m)}\right\|^2.$$

Remarquons que minimiser $\sum_{i=1}^{n} f_i(\boldsymbol{\theta})$ revient à minimiser le critère (3). On montre alors qu'il existe $L = L_\epsilon > 0$ tel que, pour $\delta_n = (k_n/\rho + n\rho)^{1/2}$, on a :

$$P\left[\inf_{|\boldsymbol{\theta}|\geq L\delta_n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta}) > \sum_{i=1}^{n} f_i\left(\sqrt{n}\widehat{\mathbf{C}}_{\boldsymbol{\rho}}^{1/2}\widehat{\boldsymbol{\theta}} - \sqrt{n}\widehat{\mathbf{C}}_{\boldsymbol{\rho}}^{1/2}\boldsymbol{\theta}^{\star}\right)|X_1,\ldots,X_n\right] > 1-\epsilon,$$

en utilisant notamment la convexité de la fonction f_i. On obtient finalement la relation

$$\mathbb{E}\left(\left\|\widehat{\Psi}_\alpha - \Psi_\alpha^{\star}\right\|_n^2 + \rho\left\|(\widehat{\Psi}_\alpha - \Psi_\alpha^{\star})^{(m)}\right\|^2 |X_1,\ldots,X_n\right) = O_P\left(\frac{1}{k_n^{2p}} + \frac{k_n}{n\rho} + \rho + \rho k_n^{2(m-p)}\right),$$

où $\|.\|_n$ est la version empirique de la semi-norme $\|.\|_2$, c'est-à-dire dans laquelle Γ est remplacé par l'opérateur de covariance empirique. Le résultat (ii) s'obtient finalement en montrant l'équivalence des deux normes sur un ensemble auquel appartient la solution du problème de minimisation.

Notons que, sous les hypothèses du théorème, si on prend $k_n \sim n^{1/(4p+1)}$ et $\gamma = 1/(4p+1)$, on obtient la vitesse de convergence $\| \widehat{\Psi}_\alpha - \Psi_\alpha \|_2^2 = O_P\left(n^{-2p/(4p+1)}\right)$. C'est la même vitesse que celle obtenue par Cardot *et. al.* [1] pour l'estimateur spline de la moyenne conditionnelle dans le cas du modèle linéaire fonctionnel. Le problème de l'optimalité de ces vitesses reste une question ouverte dans le cadre des modèles linéaires fonctionnels. On peut remarquer qu'elles dépendent de manière importante du comportement de la plus petite valeur propre de $\widehat{C}_\rho$. Nos travaux se poursuivent dans cette direction.

Remerciements. Nous remercions les participants au groupe de travail "STAPH" de Toulouse ainsi que les rapporteurs de cette note dont les commentaires ont contribué à améliorer ce travail.

Références

[1] Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline Estimators for the Functional Linear Model. *Statistica Sinica*, **13**, 571-591.

[2] de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New-York.

[3] Ferraty, F. et Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, **17**, 545-564.

[4] He, X. et Shi, P. (1994). Convergence Rate of B-Spline Estimators of Nonparametric Conditional Quantile Functions. *Nonparametric Statistics*, **3**, 299-308.

[5] Koenker, R. et Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.

[6] Lejeune, M. et Sarda, P. (1988). Quantile Regression : A Nonparametric Approch. *Computational Statistics and Data Analysis*, **6**, 229-239.

[7] Ramsay, J.O. et Dalzell, C.J. (1991). Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, **3**, 539-572.

[8] Ramsay, J.O. et Silverman, B.W. (1997). *Functional Data Analysis*. Springer-Verlag.